Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media

Isabel Cachola[*‡] **Eric Holgate**[*†] **Daniel Preoţiuc-Pietro**[◇] **Junyi Jessy Li**[†]
[‡]Department of Mathematics, [†]Department of Linguistics,
The University of Texas at Austin
{isabelcachola@,holgate@,jessy@austin.}utexas.edu
[◇]Computer and Information Science, University of Pennsylvania
danielpr@sas.upenn.edu

Abstract

Vulgarity is a common linguistic expression and is used to perform several linguistic functions. Understanding their usage can aid both linguistic and psychological phenomena as well as benefit downstream natural language processing applications such as sentiment analysis. This study performs a large-scale, data-driven empirical analysis of vulgar words using social media data. We analyze the socio-cultural and pragmatic aspects of vulgarity using tweets from users with known demographics. Further, we collect sentiment ratings for vulgar tweets to study the relationship between the use of vulgar words and perceived sentiment and show that explicitly modeling vulgar words can boost sentiment analysis performance.

1 Introduction

Vulgarity is common in language use, with vulgar words appearing with a frequency estimated between 0.5% and 0.7% in day-to-day conversational speech (Jay, 2009; Mehl et al., 2007) and 1.15% in Twitter (Wang et al., 2014). Vulgarity can employed for multiple goals: as an intensifier for subjective opinions, as a way to offend or express hate speech towards others, to describe vulgar activities or as a way to signal an informal conversation. Understanding the expression of vulgarity in naturally occurring text is thus important for several disciplines such as linguistics, which aims to better understand the pragmatics of vulgarity, for computer scientists which can explicitly model vulgarity in downstream NLP applications and for psychologists who study the socio-cultural factors of profanity.

Social media offers researchers access to vast volumes of naturally occurring user-generated content, with language use on social media containing a high level of expression of thoughts, opinions and emotions (Java et al., 2007; Kouloumpis et al., 2011). Furthermore, it is a platform for observing written interactions and conversations between users (Ritter et al., 2010). Thus, social media is a ideal medium to observe and study the expression of vulgar words and, in addition, allows us to study the socio-cultural context of this phenomenon.

Given the fact that most thoughts can be rephrased to not include vulgarity, the use of vulgar words indicates a purposeful attempt of performing a specific function. Table 1 includes examples of tweets containing vulgar words that perform different functions.

Vulgarity is often employed used to express emotion in language and can be used either to express negative sentiment or emotions or to intensify the sentiment present in the tweet (Wang, 2013). In one of the examples, 'I am stupid as f*ck' conveys more intense anger, while 'I am stupid' conveys a less emotional expression of irritation. Hence, understanding vulgar words is expected to have practical implications in sentiment analysis on social media. Furthermore, vulgar word usage is dependent on the user socio-cultural context with demographic and social information shown to improve sentiment analysis performance (Volkova et al., 2013; Yang and Eisenstein, 2017).

Hence, this study aims to address the following research questions:

- Is the expression of vulgarity and its function different across author demographic traits?

[*] Equal contribution.

Proceedings of the 27th International Conference on Computational Linguistics, pages 2927–2938
Santa Fe, New Mexico, USA, August 20-26, 2018.

Tweet	Function
I have made the awful decision to stay up all night. Why? Because I'm stupid as fuck.	emphasize
It's annoying when a program does the "you sure you want to quit?" thing but I'd be lying if I said it hasn't saved my ass a number of times	neutral/idiomatic
U got some shitty weather Illinois.... I miss texas heat	express negative sentiment
<USER> I had that game it was the shit I believe it was something like baseball 2121	express positive sentiment

Table 1: Examples of tweets with vulgar words and their function.

- Does vulgarity impact perception of sentiment?
- Does modeling vulgarity explicitly help sentiment prediction?

To this end, we collect a new data set of 6.8K tweets labeled for sentiment on a five-point scale by nine annotators. We found that the expression of vulgarity interacts with many key demographic variables, including gender, age, education, income, religiosity and political ideology. Vulgarity is also used to convey different sentiment polarities by users that differ in age and religiosity. When studying the impact of vulgarity on sentiment perception, we show that vulgarity is most often used as an intensifier for the original sentiment. Finally, we show that explicitly informing sentiment analysis systems about vulgarity usage helps boost predictive performance.

This study aims to develop a quantitative understanding of the way in which social demographics interact with how vulgarity is actively and intentionally employed and show that this can be used to improve end-user applications such as sentiment analysis. We release our novel annotated corpus as well as the accompanying code to the community at `https://github.com/ericholgate/vulgartwitter`.

2 Related Work

Vulgarity The study of vulgar language – also referred to as profanity or use of swear/curse words despite the fact that they can be used for different goals – is of interest to linguists, psychologists and computer scientists. Vulgar words are very versatile, with a vulgar word being able to perform different interpersonal functions according to different contexts (Andersson and Trudgill, 1990). Several studies have examined these different uses and pragmatic functions. Pinker (2007) groups use of swear words into five roles: abusive (intention to offend or cause psychological harm), cathartic (used in response to pain), dysphemistic (intention to convey negative sentiment), emphatic (intention to draw attention) and idiomatic (used for no purpose or to signal informality). Wang (2013) identifies four pragmatic roles of swear words: express emotion, express emphasis (either in presence of additional sentiment or not), to signal group identity and for aggression – in this latter case, swearing is a face-threatening act (Brown and Levinson, 1987). Overall, the function and effects of vulgar word usage are highly dependent on the context and the social factors (e.g., inter-personal relationships, setting), which we aim to further study and capitalize on in this study.

Recently, there has been a surge in interest in studying hate speech online (Warner and Hirschberg, 2012). A better understanding of vulgar words and the context they are used in can aid models that aim to automatically detect and categorize hate speech, which is harder to distinguish when profanity is employed (Malmasi and Zampieri, 2018).

Vulgar words have been quantitatively studied in social media and online communities. Liu et al. (2010) identified and filtered vulgar content distributed via peer-to-peer applications. Wang et al. (2014) presents a quantitative analysis of the frequency of curse word usage on Twitter, their variation with time and location, their gender usage and use in conversations across genders. An analysis of swear word usage across gender and age on Twitter was performed in Gauthier et al. (2015). Vulgar words are obfuscated in mediums where messages are censored or for other goals, and an analysis of these techniques is presented in Laboreiro and Oliveira (2014). Our paper expands the scope of this research by studying several other demographic traits beyond gender, aiming to identify different uses of swear

words through eliciting sentiment annotations and uses these to improve sentiment prediction methods.

Sentiment analysis Sentiment analysis on Twitter is a very popular research topic, anchored in a wide range of industry applications and with several shared tasks over the past years (Rosenthal et al., 2017). Several studies have looked at the effects of intensifiers and words that switch their polarity in sentiment dictionaries (Flekova et al., 2015). Vulgar words are commonly part of popular sentiment and emotion analysis lexicons (Mohammad and Turney, 2013).

Demographics & Vulgarity Use of vulgar or swear words is influenced by pragmatic or contextual factors such as the inter-personal relationship between the speakers or their social factors such as gender, occupation or social status (Jay and Janschewitz, 2008). The most studied social factor was gender, with several studies finding that males employ profanity much more often than females (Selnow, 1985; Wang et al., 2014). Other social factors such as age, religiosity or social status were found to be related with the rate of using vulgar words (McEnery, 2004). In studying the task of user trait prediction from social media texts, several vulgar words were shown to be characteristic of demographic traits such as political ideology (Preoţiuc-Pietro et al., 2017).

Explicitly modeling demographic (Lynn et al., 2017) and other social factors such as community membership (Yang and Eisenstein, 2017) have to date been successfully used for improving accuracy of several tasks including sentiment analysis (Volkova et al., 2013) or document classification (Hovy, 2015).

3 Data

In exploring the role of vulgarity in written interactions, social media and Twitter in particular stands out as an ideal medium for collecting data. Twitter provides vast volumes of naturally occurring text, which are less affected by formal and curated text such as newswire, and can be studied together with socio-demographic attributes of their authors.

Data Collection We collect a novel corpus of tweets, all of which contain vulgar words, and annotate this for sentiment, as to the best of our knowledge, no such corpus focusing on vulgar expressions exists in past research. This data is used in the sentiment analysis section of this study.

In order to enable the analysis of demographics and vulgarity in the analysis section of this study, the above tweets were deliberately sampled from those posted by a set of 4,132 users with known socio-demographic traits obtained through asking the users to self-report them in an online survey. For the demographic analysis, we have downloaded and used all most recent 3,200 tweets (per Twitter API limitations) from these users. This data set was used in prior research and for a full description of the data collection process and quality control processes, we refer the interested reader to Preoţiuc-Pietro et al. (2017).

Demographic Variables and Coding The Twitter users self-reported through a survey the following demographic variables: gender, age, education level, annual income level, faith and political ideology. All users solicited for data collection were from the United States in order to limit cultural variation.

- Gender was considered as binary[2] and coded with Female – 1 and Male – 0. All other variables are treated as ordinal variables.
- Age is represented as a integer value in the 13–90 year old interval.
- Education level is coded as an ordinal variable with 6 values representing the highest degree obtained, with the lowest being 'No high school degree' (coded as 1) and the highest being 'Advanced Degree (e.g. PhD)' (coded as 6).
- Income level is coded as on ordinal variable with 8 values representing the annual income of the person, randing from '<20,000$' to '>200,000$').
- Faith is coded as a ordinal variable with 6 levels, representing the average number a time a user attends religious service. The answers range from 'Never' (coded as 1) through to 'Multiple times a week' (coded as 6).

[2]We asked users to report gender as either 'Female', 'Male' or an open-ended field, and removed the few users which did not select 'Male' or 'Female'

- Political ideology is measured on the conservative–liberal spectrum, the common way of representing ideology in the US (Ellis and Stimson, 2012). Ideology option are: Very conservative (1), Conservative (2), Moderately conservative (3), Moderate (4), Moderately liberal (5), Liberal (6), Very liberal (7); Other (8) and Apathetic (9). In order to convert this to a ordinal scale, we only perform political ideology analysis with users with a ideology in the set 1, 2, 3, 5, 6, 7 (1290 users removed in total, similar to (Preoţiuc-Pietro et al., 2017)).

Identifying vulgar tweets In order to identify tweets containing vulgarity, we start with the vulgarity lexicon available at `www.noswearing.com`. This lexicon comprises 349 items including general vulgarities, slurs, and sex-related terms.[3] We manually removed a total of 82 items from this list that were deemed not to be unambiguously vulgar.[4] Additionally, we employed a manually crafted list of regular expressions to identify common intentional spelling variations in vulgar terms (e.g., vowel reduplication such as *fuuuuu*k* or sibilant reduplication such as *asssss*).

We identified 261,592 tweets containing at least one vulgar word, from 3,626 users featuring 222 unique vulgar tokens, many of which are either compounds of vulgar tokens with non-vulgar words or are morphological derivations (e.g., *assbag* or *f*ck* vs. *f*cking*). Of these, 149 words occur in the data less than 100 times, while the top 20 most frequent words account for 90% of total vulgar word occurrences. The median frequency of the vulgarity lexicon is 27 in our larger data set. These results are in line with previous analyses of vulgar word frequencies and their distribution (Wang et al., 2014).

Data Processing To preserve anonymity, we replace URL's by a <URL> token and usernames are masked by a <USER> token. Further, punctuation is removed and all words are lowercased.

Sentiment Annotation Human annotations of sentiment were solicited for 7,800 vulgar tweets (excluding retweets) via Amazon Mechanical Turk (MTurk) platform. These tweets were sampled uniformly from a randomly sampled set of 2,000 users in our corpus that posted at least one vulgar tweet. We perform this sampling of users in order to have a broad range of users, but have more than one sample tweet for each user.

The task setup and guidelines follow the settings of SemEval 2016 Task 4 subtask C (Nakov et al., 2016). Annotators evaluated tweet sentiment on a five-point scale: (1) *very negative*, (2) *somewhat negative*, (3) *neutral*, (4) *somewhat positive*, and (5) *very positive*. The numeric coding for this scale follows a gradual movement from more negative to more positive, allowing sentiment to be treated as a continuous variable. A sixth option, *not applicable*, was available to workers to cover instances in which there was missing contextual information or a tweet was illegible. Each tweet was annotated by nine different and independent annotators to allow for sufficient data to generalize across individual sentiment perception boundaries.[5]

Annotation results were checked for inter-annotator agreement by comparing individual user responses to the average annotation of their peers across all tweets that they annotated (Li et al., 2016). Annotations from users with a Spearman correlation coefficient less than 0.3 were removed from computing consensus labels. Tweets with less than five remaining annotations were excluded. We also excluded tweets whose majority ratings are *not applicable*. This resulted in a final data set consisting of 6,791 annotated vulgar tweets. Following the procedure for SemEval, Sentiment annotations for these tweets were consolidated by majority vote. If a majority rating was not present, the consolidated score was set to the mean of all ratings (ratings of *not applicable* were excluded).

The consolidated sentiment ratings were distributed as follows: 1.94% *very positive*, 14.57% *somewhat positive*, 26.51% *neutral*, 46.62% *somewhat negative*, 10.14% *very negative*.

[3]While this initial method we use is not fully accurate (Sood et al., 2012a), it reaches high accuracy levels of more than 90% (Sood et al., 2012b)

[4]These terms were largely anatomical words or general verbs like *penis*, *vagina*, and *blow*, but some identity descriptors like *gay*, *queer*, and *lesbian* were also excluded after manual review of a large sampling of uses revealed they were not overwhelmingly employed as slurs.

[5]We asserted the following qualifications on MTurk: locale=US, approval rate >90%, number of HITs approved >100.

Demographic Trait	Pearson r	p-value
Gender	-0.077	1.61^{-4}
Age	-0.233	6.64^{-31}
Education	-0.100	7.62^{-07}
Income	-0.087	1.73^{-05}
Faith	-0.187	2.74^{-20}
Political Ideology	0.124	8.69^{-10}

Table 2: Relationship between user demographic traits and percentage of vulgar tweets of total tweets sent by the user. Results are computed using Pearson correlation (point-biserial correlation for gender), with gender and age as controls in partial correlation.

4 Analysis

4.1 Demographics & Frequency of Vulgarity

We first analyze the differences in frequency of employing vulgarity and how this relates to demographic traits. We use partial Pearson correlation where the dependent variable is the fraction of vulgar tweets from the total number of tweets sent by a user. For all analyses, we consider gender and age basic traits and control for data skew by introducing both variables as controls in partial correlation, as done in prior work (Schwartz et al., 2013; Preoţiuc-Pietro et al., 2016). When studying age and gender, we use only the other trait as the control. We have experimented with percentage of vulgar words as an alternative outcome, but the results were very similar, hence we exclude them for brevity. We also experimented with log-scaling the age variable, but found no major differences.

Gender results show that females are less likely to post vulgar tweets than males. Several previous studies have reported a similar effect, in written text in the BNC (McEnery, 2004), on social networks (Wang et al., 2014), speech (Jay, 1980; Jay, 1992) or as self-reported using questionnaires (Selnow, 1985; Pilotti et al., 2012). In our data, the average vulgar tweet percentages per user are 3.332% for males and 3.060% for females.

Age exhibits the largest effect on posting vulgar tweets, with younger users much more likely than older to post, which aligns to analysis of speech (Jay, 1992) and tweets as previously performed in Gauthier et al. (2015).

Both higher education and income are anti-correlated with usage of vulgarity on social media even when controlling for gender, with education being slightly more strongly associated with lack of vulgarity than income. Previously, McEnery (2004) suggested that social rank, which is related to both education and income, is anti-correlated to use of swear words. In a study on perceptions of user traits from tweets, raters that were exposed to vulgar words were more likely to consider the user as being less educated than in reality (Carpenter et al., 2016).

Faith is anti-correlated with use of vulgar words with the second strongest effect, a correlation also previously identified by Jay (1992).

Finally, liberal users are more likely to use vulgarity on social media, an association on Twitter also uncovered by Sylwester and Purver (2015) and Preoţiuc-Pietro et al. (2017).

4.2 Demographics & Sentiment Perception

Next, we analyze the relation between the demographic traits of the tweet authors to the perceived (annotated) sentiment of the tweet. We perform this analysis in order to uncover potentially different types of uses of vulgar words. If vulgar words are used as an intensifier, they are used with both positive and negative sentiment, while if expressing emotion they are expressed mainly with negative sentiment. If vulgar words are used in neutral tweets, then their role is more likely to be idiomatic.

We test this using partial Pearson correlation with user demographics as the independent variable and with the polarity of the rating (positive, negative or neutral) as a dependent indicator variable. Again, we consider gender and age as basic demographic traits and introduce these variables as controls. Similarly, for age and gender analysis, we the other basic trait as the only control. The results of these analyses are

	Positive vs. All		Negative vs. All		Neutral vs. All	
Demographic Trait	r	p	r	p	r	p
Gender	0.024	0.334	-0.006	0.797	-0.030	0.222
Age	-0.107	$\mathbf{2.11}^{-5}$	0.061	**0.015**	0.093	$\mathbf{2.305}^{-4}$
Education	-0.049	0.052	-0.018	0.470	0.076	**0.002**
Income	-0.024	0.324	0.016	0.524	0.001	0.966
Faith	-0.108	$\mathbf{1.640}^{-5}$	0.008	0.741	0.128	$\mathbf{3.539}^{-7}$
Political Ideology	-0.032	0.276	0.057	0.051	0.008	0.765

Table 3: Relationship between user demographic traits and perceived sentiment rating by MTurk workers. Results are computed once more using Pearson correlation (point-biserial correlation for gender), with gender and age as controls in partial correlation.

presented in Table 3.

Positive Sentiment With positive sentiment as an indicator variable, we see that both age and faith have significant correlations. This indicates that younger and less religious users are more likely to express vulgar words in presence of positive sentiment, which represents use of vulgarity as an intensifier for sentiment.

Negative Sentiment With negative sentiment as an indicator variable, we again see a weaker (yet still significant) correlation with age. This shows that older users are prefer to use vulgar words to deliberately express negative emotions and sentiment, with other functions being less frequent.

Neutral Sentiment Finally, vulgar words are present with neutral sentiment when this is used idiomatically for signaling informality or other functions. Neutral sentiment is most often employed by religious, older and more educated users. The latter correlation is especially relevant, as higher education leads to better social adaptation.

Overall, gender, income and political ideology exhibit no significant correlations with sentiment ratings of vulgar tweets.

4.3 Vulgar vs. Censored Tweet Sentiment Perception

To test whether vulgarity impacts the *perceived* sentiment of tweets, we removed the vulgar terms in the tweets from the vulgar corpus, arriving at an *original-censored* parallel dataset. We then crowdsourced the sentiment judgments of the censored tweets using the same method as before.

Among the censored tweets, annotators indicated that 355 —around 5%—of the tweets were rated by more than 5 (out of 9) workers to be unintelligible after censorship. The distribution of sentiment in this subset was originally 21.40% positive, 21.97% neutral, 56.62% negative. This loss of intelligibility is expected and an additional indication that vulgarity bears semantic content and by censoring these words, the resulting tweet looses coherence.

To determine the extent to which the use of vulgar terms impacts the perception of tweet sentiment, we calculated the direction and magnitude of the change in sentiment between the vulgar and censored pairs. For original tweets that are not neutral, we consider three situations: (a) **intensify**, where the magnitude of the sentiment towards a tweet intensifies after censorship, e.g. with score changes from 2 to 1 (for negative tweets) or from 4 to 5 (for positive tweets); (b) **weaken**, where the magnitude changes in the other direction, e.g., with score changes from 1 to 2 (for negative tweets) and from 5 to 4 (for positive tweets); (c) **flip**, where the censored sentiment is opposite from the original sentiment, i.e., negative becomes positive or positive becomes negative. For neutral tweets, we consider situations where the censored sentiment becomes positive (**to-pos**), or negative (**to-neg**), or stays neutral (**same**).

In Table 4, we present the numbers and magnitudes of changes observed in these 6,436 tweets (this number reflects the size of the censored corpus after excluding the 593 which were rated unintelligible and those which did not have at least five ratings after removing annotators with low inter-annotator agreement scores). Clearly, for original positive or negative tweets, censoring has a significant weakening

	Positive			Negative			Neutral		
	intensify	weaken	flip	intensify	weaken	flip	to-pos	to-neg	same
% changed	15.17	34.09	3.69	5.53	40.34	7.44	24.07	19.34	56.59
mean $\|\Delta\|$	0.591	0.785	1.597	0.521	0.936	1.569	0.817	0.825	—

Table 4: Change in perceived sentiment score with censored vulgarity. Results are grouped by initial perceived sentiment polarity, scored from 1 (*very negative*) to 5 (*very positive*). The *intensify* subcondition indicates that the censored tweet had the same polarity, but with greater magnitude (i.e., original 2, censored 1; or original 4, censored 5). The *weaken* subcondition describes the opposite effect. The *flip* subcondition indicates that the polarity of the perceived sentiment inverted. All of the changes are statistically significant ($p < 0.01$)

effect for sentiment. Censored tweets that are originally negative also have a stronger tendency than originally positive ones to flip sentiment. For neutral tweets, most of the time the sentiment does not change.

Kruskal-Wallis H-tests were conducted to determine if changes in sentiment rating were significant, and all results were shown to be statistically significant ($p < 0.01$). The strongest effect was seen in the *to-pos* and *negative-flip* conditions, indicating that vulgarity is frequently employed by users to emphasize or introduce a negative sentiment.

About half of the tweets in the corpus did not exhibit a change in sentiment. Of those that didn't change, 21.31% were positive, 21.11% were neutral, and 56.14% were negative. Vulgar tweets which were initially positive or negative were slightly more likely to exhibit some change in perceived sentiment rating than not, while vulgar tweets that were initially neutral were 8% more likel y to exhibit no change at all than they are to exhibit change.

5 Modeling Vulgarity

We now consider whether explicitly inserting knowledge about vulgarity into sentiment models will help sentiment prediction. To do this, we build on top of a base model that has been shown to be effective in Twitter sentiment prediction (Rosenthal et al., 2017; Cliche, 2017), and propose three ways of inserting vulgar features into the system.

Base architecture The basis of our system is a bidirectional Long-Short Term Memory Network (LSTM) (Hochreiter and Schmidhuber, 1997), which utilizes memory cells to capture long-term dependencies. This architecture is similar to the LSTM models in BB_twtr (Cliche, 2017), the best system participated in SemEval 2017 task 4 subtask C (Rosenthal et al., 2017).

For a given tweet of n words, we encode each word w_t into an embedding vector x_t; we then pass the embedding matrix to a bidirectional LSTM, arriving at a sentence representation where each hidden state h_t for word w_t encodes the context from both words before w_t (i.e., w_1 to w_t) and after w_t (i.e., w_t to w_n). The hidden state is the concatenation of two directions: $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$.

The bidirectional LSTM sentence representation is then passed into a dense layer with a RELU activation function $RELU(x) = max(0, x)$. The dense layer is then followed by a dropout layer, and a softmax classification layer.

Vulgarity features We propose and compare three different ways to account for vulgarity:

- **Masking**: Mask all vulgar words with a vulgar token, e.g., *This is the <VG>*. This method essentially considers all vulgar words to be the same token and informs the model of such, but removes lexical variances and different expressions of vulgarity.
- **Token insertion**: Insert a vulgar token after each vulgar word, e.g, *This is the shit <VG>*. This method allows the curse word to retain its original meaning, while giving it the same representation as other curse words.

System	Overall	Very positive	Positive	Neutral	Negative	Very negative
Bi-LSTM	0.791	3.000	1.978	1.008	0.049	1.084
+ Masking	0.898	**2.000**	**1.030**	**0.028**	1.000	1.979
+ Token Insertion	0.761	3.000	2.000	0.996	**0.002**	**1.000**
+ Concatenation	**0.759**	3.000	1.993	0.992	**0.002**	**1.000**

Table 5: MAE for each system on Vulgar-Tweet dataset. The best per-class results are significantly better than the baseline Bi-LSTM with the Wilcoxon signed-rank test.

- **Concatenation**: Count the number of vulgar words n_v in a tweet, and use it as a feature in concatenation with the tweet representation from Bi-LSTM h_t, before feeding it to the next layer, whose new input would be $[h_t, n_v]$. Effectively, this informs the classifier of the number of vulgar words, but removes contextual information of where a curse word appears.

6 Experiments

The goal of our evaluation is to assess how explicit knowledge of vulgarity can help in sentiment prediction. Hence we focus on tweets with vulgarity present, and we train and evaluate on our Vulgar-Tweet dataset. We further compare the aforementioned three ways to encode vulgarity features.

6.1 Systems and training

Systems We train and compare the following systems: *Bi-LSTM*, i.e., the base architecture without explicit knowledge about vulgarity; and the base architecture with each of the vulgarity features stated: *Bi-LSTM + Masking*, *Bi-LSTM + Token insertion*, and *Bi-LSTM + Concatenation*.

Data We develop, train and evaluate our system on our Vulgar-Tweet dataset, from which we carved out a development set of 500 tweets, a test set of 1,000 tweets. The remaining 5,291 tweets were used for training.

It is worth noting that the SemEval data is not viable for evaluating the effects of modeling vulgarity. There are only 1,235 (4%) of vulgar tweets in the 2016 data, and 339 (2.75%) in the 2017 testing data. For the sake of comparison, we also train a system with SemEval 2016 training data, and report on the 2017 test set where there is at least one vulgar word in each tweet.

Settings We used 200 dimensional embeddings trained on a corpus of 50 million tweets (Astudillo et al., 2015). We do not tune the embeddings. Embeddings for unknown words are randomly initialized. The LSTM dimension is 128 and we use a batch size of 256. We use accuracy as our metric and optimize using the Adam optimizer (Kingma and Ba, 2014). The dropout layer uses a dropout rate of 0.4. We train until the validation accuracy stops improving with a learning rate of 0.001. All hyperparameters are tuned on the SemEval development set. The model is implemented in Keras (Chollet and others, 2015).

6.2 Evaluation

Metrics Since our data is annotated on a 5-point ordinal scale, we used mean absolute error (MAE), which asserts more penalty when the predicted label is further away from the true label, i.e., if the system predicts 1, and the true label is -2, the error will be 3 (instead of just "incorrect"). On a test set Te where the true class of tweet i is c_i, the mean absolute error (MAE) is:

$$MAE = \frac{1}{|Te|} \sum_{x_i \in Te} |y_{pred} - y_{true}| \quad (1)$$

We report the overall MAE as well as the MAE values for each sentiment class.

Results Table 5 tabulates the MAE values for each system. Due to the fact that our data set is largely negative, our systems in general are better at predicting negative and very negative tweets. Except masking, vulgarity features improves the overall performance of the system. Concatenating a vulgar token

Bi-LSTM	+ Masking	+ Token Insertion	+ Concatenation
1.320	1.666	**1.068**	1.148

Table 6: MAE on SemEval 2017 task 4 subtask C test set, vulgar tweets only.

Text	**True Label**	**Baseline Prediction**	**Insertion Prediction**	**Concatenation Prediction**
the bitch is back jamies new tumblr page	Negative	Neutral	Negative	Negative
welcome to my personal hell	Negative	Neutral	Negative	Negative
the simpsons donut hell apptrailers 0 likes	Neutral	Neutral	Negative	Negative
so fucking excited	Very Positive	Neutral	Negative	Negative

Table 7: Example Tweets and their predictions.

produced the lowest MAE overall, followed by insertion. The best per-class results are significantly better than the baseline Bi-LSTM with the Wilcoxon signed-rank test.

Across sentiment labels, masking improves the prediction of positive tweets; however, the prediction of negative tweets suffer. With masking, the actual vulgar word is replaced by a special token, stripping its meaning. One reason for this behavior could be that for positive tweets, vulgarity is more often used as an intensifier than other sentiment classes (Table 4, positive-weaken), so there is less of an impact on the loss of meaning, vs. the benefit of explicit vulgarity information. Token insertion and concatenation improves the prediction of negative tweets, and the prediction of non-negative tweets remain stable. This shows that retaining the meaning of vulgar expressions is in general more helpful and especially for negative sentiment.

Finally, we report on systems trained on SemEval and tested on the 339 vulgar-only tweets in the SemEval test set, shown in Table 6. Again, we see that inserting and concatenating vulgar word count both improves performance.[6]

Qualitative Analysis In Table 7, we show example tweets along with their labels and predictions. The first two are correct predictions, and the two others are incorrect predictions.

In the first tweet, "is back" is often used in positive sentences. For example, the tweet "pretty little liars is back in action hell yes" is rated very positive. However, "bitch" is often used as a negative word, making the baseline predict the sentence is neutral. The addition of vulgar features gave greater significance to the word "bitch," pulling the sentiment in the negative direction. The baseline may have rated the second tweet as neutral because the combination of the positive word "welcome" and the negative word "hell" gives the tweet a neutral prediction. Again, the vulgarity features give "hell" more significance, allowing the system to pick up on the negativity of the tweet.

In the third sentence, the presence of the word "hell" pushed the prediction to be negative, although in this case it made the prediction incorrect. In the last sentence, the systems with vulgar features predicted negative although the sentence is very positive. The baseline predicted negative, indicating that it did not have enough power to counterbalance "fucking" and "excited". However, our systems pulled the sentiment in the wrong direction. This may also be because vulgarity is often used in a negative form in this context (e.g. "so fucking angry").

7 Conclusions

We have introduced a new data set of 6.8K vulgar tweets labeled for sentiment on a five-point scale by nine annotators. Analysis of the ratings revealed that the expressiveness of vulgarity interacts with a number of demographic features, including age, gender, education, income, religiosity, and political ideology. Vulgarity is used to different ends by users who differ in age and faith. An examination of the impact of vulgarity on tweet sentiment perception showed that, in cases where the presence of vulgarity

[6]The MAE of Bi-LSTM on all the SemEval 2017 test set is 0.52, comparable with submitted systems (Rosenthal et al., 2017).

influences sentiment perception, it is most often used to intensify existing sentiment. Finally, we have shown that utilizing vulgarity-centric features yields increased sentiment analysis system performance. Future work will look directly at modeling the functions of vulgar words.

References

Lars-Gunnar Andersson and Peter Trudgill. 1990. *Bad language*. Penguin.

Ramón Astudillo, Silvio Amir, Wang Ling, Mário Silva, and Isabel Trancoso. 2015. Learning word representations from scarce and noisy data with embedding subspaces. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1074–1084.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some Universals in Language Usage*, volume 4. Cambridge University Press.

Jordan Carpenter, Daniel Preoţiuc-Pietro, Lucie Flekova, Salvatore Giorgi, Courtney Hagan, Margaret Kern, Anneke Buffone, Lyle Ungar, and Martin Seligman. 2016. Real Men don't say 'cute': Using Automatic Language Analysis to Isolate Inaccurate Aspects of Stereotypes. *Social Psychological and Personality Science*, 8.

François Chollet et al. 2015. Keras. `https://github.com/keras-team/keras`.

Mathieu Cliche. 2017. BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 573–580.

Christopher Ellis and James A Stimson. 2012. *Ideology in America*. Cambridge University Press.

Lucie Flekova, Eugen Ruppert, and Daniel Preoţiuc-Pietro. 2015. Analysing domain suitability of a sentiment lexicon by identifying distributionally bipolar words. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*.

Michael Gauthier, Adrien Guille, A Deseille, and Fabien Rico. 2015. Text mining and twitter to analyze british swearing habits. *Handbook of Twitter for Research*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 752–762.

Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65.

Timothy Jay and Kristin Janschewitz. 2008. The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture*, 4(2):267–288.

Timothy B Jay. 1980. Sex roles and dirty word usage: A review of the literature and a reply to Haas. *Psychological Bulletin*, 88:614–621.

Timothy Jay. 1992. *Cursing in America: A Psycholinguistic Study of Dirty Language in the Courts, in the Movies, in the Schoolyards, and on the Streets*. John Benjamins Publishing.

Timothy Jay. 2009. The utility and ubiquity of taboo words. *Perspectives on Psychological Science*, 4(2):153–161.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 538–541.

Gustavo Laboreiro and Eugénio Oliveira. 2014. What we can learn from looking at profanity. In *International Conference on Computational Processing of the Portuguese Language*, pages 108–113.

Junyi Jessy Li, Bridget O'Daniel, Yi Wu, Wenli Zhao, and Ani Nenkova. 2016. Improving the annotation of sentence specificity. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*.

Xiangtao Liu, Xueqi Cheng, Jingyuan Li, Haijun Zhai, and Shuo Bai. 2010. Identifying vulgar content in emule network through text classification. In *IEEE International Conference on Intelligence and Security Informatics*, pages 168–168.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H Andrew Schwartz. 2017. Human centered nlp with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202.

Tony McEnery. 2004. *Swearing in English: Bad language, purity and power from 1586 to the present*. Routledge.

Matthias R Mehl, Simine Vazire, Nairán Ramírez-Esparza, Richard B Slatcher, and James W Pennebaker. 2007. Are women really more talkative than men? *Science*, 317(5834):82–82.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 1–18.

Maura Pilotti, Jennifer Almand, Salif Mahamane, and Melanie Martinez. 2012. Taboo words in expressive language: Do sex and primary language matter. *American International Journal of Contemporary Research*, 2(2):17–26.

Steven Pinker. 2007. *The stuff of thought: Language as a window into human nature*. Penguin.

Daniel Preoţiuc-Pietro, Jordan Carpenter, Salvatore Giorgi, and Lyle Ungar. 2016. Studying the Dark Triad of Personality using Twitter Behavior. In *Proceedings of the 25th ACM Conference on Information and Knowledge Management*, CIKM, pages 761–770.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 729–740.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 502–518.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-vocabulary Approach. *PLoS ONE*, 8(9):1–16.

Gary W Selnow. 1985. Sex differences in uses and perceptions of profanity. *Sex Roles*, 12(3-4):303–312.

Sara Sood, Judd Antin, and Elizabeth Churchill. 2012a. Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1481–1490.

Sara Owsley Sood, Judd Antin, and Elizabeth F Churchill. 2012b. Using crowdsourcing to improve profanity detection. In *AAAI Spring Symposium: Wisdom of the Crowd*, volume 12, page 06.

Karolina Sylwester and Matthew Purver. 2015. Twitter Language Use Reflects Psychological Differences between Democrats and Republicans. *PLoS ONE*, 10(9), 09.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2014. Cursing in English on Twitter. In *Proceedings of the 17th ACM conference on Computer Supported Cooperative Work & Social Computing*, pages 415–425.

Na Wang. 2013. An analysis of the pragmatic functions of swearing in interpersonal talk. *Griffith Working Papers in Pragmatics and Intercultural Communication*, 6:71–79.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26.

Yi Yang and Jacob Eisenstein. 2017. Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association for Computational Linguistics*, pages 295–307.